Digital Inclusion in Nothern England: Training Women from Underrepresented Communities in Tech: A Data Analytics Case Study

Olanrewaju Tahir Aduragba
Department of Computer Science
Durham University
Durham, UK
olanrewaju.m.aduragba@durham.ac.uk

Jialin Yu
Department of Computer Science
Durham University
Durham, UK
jialin.yu@durham.ac.uk

Prof. Alexandra I. Cristea
Department of Computer Science
Durham University
Durham, UK
alexandra.i.cristea@durham.ac.uk

Dr Mariann Hardey
Durham Business School
Durham University
Durham, UK
mariann.hardey@durham.ac.uk

Prof. Sue Black
Department of Computer Science
Durham University
Durham, UK
sue.black@durham.ac.uk

***Abstract*— The TechUPWomen programme takes 100 women from the Midlands and North of England, particularly from underrepresented communities, with degrees or experience in any subject area, retrains them in technology and upon graduation guarantees an interview with a company. The retraining programme, developed by the Partner Universities in conjunction with the Industrial Partners, has modules at level 6/7 including: Technology: coding, data science, cyber security, machine learning, agile project management; Workplace readiness skills: public speaking, clear communication, working as a team. In this paper, we introduce, for the first time, the TechUPWomen programme, and we analyse its temporal evolution and special features via a data analytics nowcasting approach. Deepening these women's experience with applied upskilling includes one-to-one mentoring (100-100), strong networking, residentials, close industry connection with two directions (non-technical & technical) and four job-focussed final tracks: business analyst, agile project manager, data scientist, developer. TechUPWomen also has significant representation of traditionally underrepresented communities, with focus on enabling instead of teaching approach. Beside the originality of the unique combination of features of the programme, this is, to the best of our knowledge, the first analysis based on data analytics of a women in tech(nology) retraining programme, based on nowcasting. Results show that the approach is effective; topic analysis shows that frequent topics include joy, BAME, networking, residential, industry, learning.**

Keywords*— *TechUPWomen, Underrepresented Communities, Digital Inclusion, Data Analytics, Computer Science Education

I. Introduction

There is a growing global, as well as UK-based, attention on narrowing the gender gap and improving participation of women and members of underrepresented groups in computer science [1]. Yet, the gender inequality in technology-related fields still very much exist [2]. According to the Higher Education Statistics Agency (HESA)[1], only about 18% of the students in higher education studying computer science were women in 2017/2018, with under 1% increase from the previous year. In contrast, the computing and information technology industry has been growing exponentially, showing an urgent national need for people in technology-related fields [3]. Due to persistent hiring challenges in the technology industry, bootcamps have sprung up to provide a fast-track entry into technology roles, while costing less in terms of tuition and time. Additionally, there is an increase in the understanding that diversity is a strength in any community, in general, and in technology, in particular [4]. Thus, underrepresented groups are particularly interesting for the technology industry. Whilst 'coding bootcamps' have appeared [5], training participants in technology roles, their effectiveness is not always clear or measurable. Especially difficult is the measuring of new features introduced during the programme, to inform further potential changes. Moreover, there has been very limited data analytics performed on the retraining programmes for women in tech(nology) during the running of the programme (as in 'nowcasting' – see more information in Section 2).

In this paper, we thus present and analyse an original programme designed for women from underrepresented communities, the TechUP programme[2]. The main research questions targeted here are thus:

RQ1. How can we use data analytics to measure retraining programmes for women in tech(nology), during the actual intervention?

RQ2. How can special intervention methods support women transition into technology roles?

The main contributions of this paper are thus to define and apply measuring methods for retraining programmes for women in technology, during the actual intervention. We also describe in the process our unique programme's features and show that they have been effective so far.

II. Background and Related Work

A. Measuring Events based on User Generated Data

Real time social media data from platforms such as Twitter, Facebook, LinkedIn, as well as query volumes from search engines are being used to track real world phenomena across a wide range of topics. Social media data generated based on large groups of users provide the potential ability to observe public opinions and activities in real time [6]. Previous research on tracking, predicting and measuring real world phenomena using social media data has been applied on tasks such as epidemiological variables [7], [8], economic

[1] https://www.hesa.ac.uk/data-and-analysis/students/what-study

[2] https://techupwomen.org/

variables such as unemployment levels [9], the demand for automobiles [10], consumer consumption metrics [11], popularities and sales of video games, music tracks and feature films [12], the happiness of Internet users as a proxy for the happiness of nations [13], and the outcomes of political races [14].

An important distinction is whether such data are being used to predict the future, or to track the present. The latter, known as '*nowcasting*', aims to utilise social and Internet derived data to quantify real-world phenomena in real-time [15], boasting previous research in present-moment happiness of nations [16], real-time mortality rates [17], influenza outbreaks [18] and voting intentions during political races [19].

In this paper we are focussing on nowcasting in the evolution and impact of the development of the TechUPWomen programme as it proceeds, using Twitter and Microsoft Teams data collected over a four-month period from July 2019 up until October 2019.

B. Topic Modelling

Topic models cluster related words into general topics or concepts and identify the topics that make up a document [20]. Topic models are popular in text processing, to analyse contents of large corpora to reveal information related to health, education and other research areas [21]–[23]. Topic models are mostly unsupervised models that do not require pre-annotations of the text documents and the topic distributions are discovered by automatically clustering words into topics and assigning those topics to documents [24]. Latent Dirichlet Allocation (LDA) [25] offers an effective approach to automatically extract topics from datasets. LDA is a powerful topic model and it has been widely applied to various text documents including social media data to identify hidden topic structure which provides further insights into data [26]. In LDA, a set of documents, D, is assumed to contain K topics that are described with by a set of words, w. Each document $d \in D$ is modelled as two probability distributions assumed to be multinomial distributions, $p(t|d)$, the probability distribution of words in document d that are currently assigned to topic t, and $p(w|t)$, the probability distribution of assignments to topic t over all documents D that come from a word w. Dirichlet prior α are assigned to the multinomial distribution θ_d over K topics, $Dir(\theta_d|\alpha)$. Likewise, for topic k, Dirichlet prior η for the multinomial distributions β_k over words is derived from $Dir(\beta_k|\eta)$ [27].

C. Network Analysis

Network analysis is a critical step for Twitter-related analyses, as it provides a general relationship between Twitter users and how they use Twitter as a tool and get involved in certain events (in our case the TechUPWomen programme). It can also reveal the social relationship between users for specific events and point towards potential links between data entities in the Twitter data network. Hence, it can be used as a tool to examine and measure how closely the users are connected and how they can be linked to the topic extracted via topic modelling (Section II.B.)

There are two main streams of network analysis in social science research [28]. One approach is to represent the connections between vertices in the graph, where single interaction between any two or more vertices is recorded separately and visualised across the whole network. A more complex network analysis defines the vertices in the network as a bigraph and includes some of the interactions as vertices in the network. This increases the complexity of the network, but it could also help explore some hidden connections between vertices. However, the disjoint vertices need to be defined by human experts, in order to incorporate the information into the network and form a bigraph. For a large-sized network, such as the Twitter social network, the former version is preferable and thus being used in this research.

Another way to represent the network analysis focuses more on the critical vertices in the graph, based on the measurement of centrality [29]. Measuring the network through centrality has been applied to many previous works on network analysis including biological network [30], sexual networks and transmission of the AIDS virus [31], organisation behaviour [32] and Twitter networks [33]. The level of scores for centrality indicates the level of frequency of visits to the vertices across the network. Hence, network pruning can be applied to an existing network. We created a network with the previous method, then pruned based on critical vertices, thus providing a lighter version of the network, using the same amount of information.

III. TechUP Programme and Proposed Analytics Methodology

A. TechUPWomen

TechUPWomen[2] is a programme funded by the Institute of Coding[3] that targets women from minority groups based in the Midlands and North of England and puts them through free online training sessions for gaining technology skills, as well as offering four residentials for motivation, catching up and networking. Similar to coding bootcamps, the programme is developed in close collaboration with industrial partners; participants are assigned one-to-one mentors and each participant is guaranteed an interview with one of the industry partners after the programme. To accommodate the diversity of needs of the women, who traditionally have multiple roles and responsibilities, our intervention programme is mostly delivered online, while providing opportunity for face-to-face meetings during the course of the program. The programme maintains a very active social media presence for participants to be engaged in collective learning processes, as well as to be exposed to relevant industry networks. Some of the social media tools used for the TechUP programme include LinkedIn, Twitter, Microsoft Teams discussions, blog and Instagram.

To understand how special intervention methods would support women transition into technology roles, in the TechUPWomen programme, our study uses such as computer science techniques, such as social media mining methods, to investigate the participants' temporal activities on social

[3] https://instituteofcoding.org

media, to measure the impact of the programme to increase women participation in computer science, to support knowledge transfer into computer science roles, as well as to test new types of knowledge exchange for women who have multiple roles and responsibilities. This includes calculating important descriptive analytics, starting with statistics of TechUPWomen-related social media posts and examining them over time. Another approach involves Natural Language Processing (NLP), such as topic modelling, to transform large corpora of social media exchanges, focus group discussion, online learning platform discussions into temporal topic trends.

B. Data Collection for Data Mining

This study analyses several data sources collected during the TechUPWomen programme. Firstly, impact-related data related to the programme was collected on Twitter using the Twitter API, based on the official programme and related Twitter hashtags. The extraction runs over the whole period of the programme, starting from July 2019 till August 2020. The data analysed in this paper reflects nowcasting up to October 2019. Likewise, participant engagement on forums, including Microsoft Teams discussions were also collected during this period. Microsoft Teams was used as a collaborative platform to share learning contents, post assignments and create a dialogue between the participants and programme support staff. The platform comprises more than 200 users, including learners, programme coordinators and mentors; out of whom 100 were active users. However, Microsoft Teams does not provide automatic extraction or export of data; thus, we had to manually extract timestamps and posts from July to October 2019 from the general team conversation.

C. Pre-processing User Generated Data

Data pre-processing techniques were applied to the data directly collected from the Twitter account, which hosted most of the event announcements for the TechUPWomen programme. The Twitter language differs from text in books and articles, and because of the text limit, texts are often shortened, and they also include distinctive uses, such as URLS, repeated letters, @ for usernames, # for hashtags and emoticons. Thus, it is important to pre-process and normalise these texts [34]. We further applied simple pre-processing techniques, such as stemming, to remove tenses and plurals from the endings of words (e.g., inspired, inspiring => inspire), and stop-word filtering to remove words that were frequent but did not contain useful information (e.g., and, the). Additionally, we used a combination of tweets' coordinates and/or user profile information, depending on which was available, to infer a location for a tweet. The average number of words after cleaning the text is 10.

D. Topic Modelling for TechUPWomen

Topic modelling was applied to examine the underlying topics in our datasets. In the current study, we considered the mixture of words in a document to originate from a set of latent topics, which came from a fixed probability distribution over the vocabulary. In the study presented in this paper, we applied LDA based on the purpose, as well as the nature of our data sources.

We used the LDA model implementation from Scikit-learn (Pedregosa *et al*, 2011) to construct our topic models. We used the default hyperparameter settings for the LDA algorithm in Scikit-learn. We used qualitative judgement to choose the number of topics as the parameter for our LDA model.

E. Network Analysis

We represent the social network interaction in TechUPWomen as a graph and use graph metrics to describe important information across the whole network. For the social network in Twitter, we remove '@' before the user account name before turning it into vertex in the network; hence we defined all the identical vertices in the network by going through all the Tweets collected. To build the complex network, whenever a vertex of a user account posted a message, which included '@' and another user account, an edge was created in the network, pointing from the first user to the next user. The metrics in the network was mostly considered as an undirected graph which means that an edge from one user to another was equivalent in the inverse direction [28]. This kind of representation of the network is more focused on the visualisation of the activity levels for each individual vertex, and as the network grows more complex, the aggregated network could reveal the most critical groups of nodes in the network, which have the most influence across the whole network.

Additionally, we focussed on the most popular nodes with the centrality method [29]. The visualisation tool we used is called 'Networkx' [36] and for measurement of centrality, the eigenvector centrality was used to measure if a given node was a hub and how that node was connected to other hubs.

IV. Results and Discussion

A. Twitter and Microsoft Teams Analysis

Figure 1 illustrates the number of engagements (including tweets, retweets and comments) from Twitter and Microsoft Teams between July and October 2019. Interestingly, the levels of engagements on both Twitter and Microsoft Teams form a similar pattern during this period. Posts before and during each residential peaked for both Twitter and Microsoft Teams. For Twitter, the highest number of posts were recorded only during the residential sessions, while on Microsoft Teams higher number of comments were also sent on other days – for instance, when assignments were due. The graph shows that the discussions were sustained on both platforms over that period of time. Whilst in terms of pattern, the two social media options show similarities, in terms of sheer numbers, the volume on Twitter is of a level of magnitude (x10) higher, in general. Also, overall, there is more variation in the communication level on Microsoft Teams, possibly also due to the lower level.

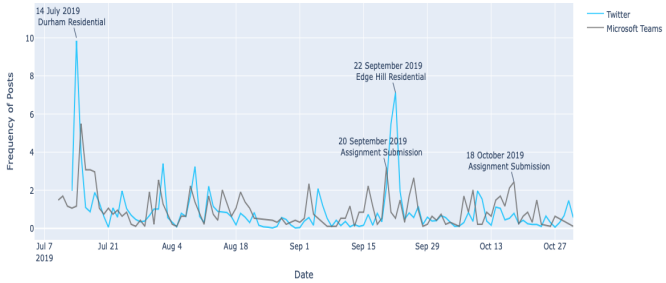

Fig. 1. Comparison of daily posts on Twitter and Microsoft Teams

Figure 2 displays the ranking of the top 20 terms, computed using the Term Frequency – Inverse Document Frequency (TF-IDF) measure, for Twitter and Microsoft Teams. Unlike term frequency, which is based on the number of occurrences, TF-IDF uses a weighting scheme based on the term frequency and the number of documents containing that term. Figure 3 shows a similar analysis in a numerical way. We can infer from the presence of "techupwomen", "womenintech" in the TF-IDF list that the participation of women in technology is being discussed on Twitter. Another set of words that are also associated with TechUP are "weekend", "residential", and "tuwres" (the latter also referring to residentials). This shows that the residentials organised by TechUPWomen generated a significant amount of discussion on Twitter. The list of frequent terms also contains words such as "great", "love", "take", "excit" (meaning 'exciting'), "inspir" (referring to 'inspirational' or similar), "amaz" ('amazing'). Overall, these observations form the basis for a more in-depth investigation of the topics discussed in our programme, which are mostly very positive, showing the diversity of our approach (such as 'colour women' or 'women', 'colour', 'disability women'). Comparing Twitter and Microsoft Teams, we can see that the latter focusses more on the 'assignment', 'think'-ing, programming languages (such as 'Python'), thus being more focussed on the 'work' performed, as opposed to generic features of the programme. Interestingly, participants are also aware (and possibly worried) about the 'time' the 'course' and 'module'(s) take.

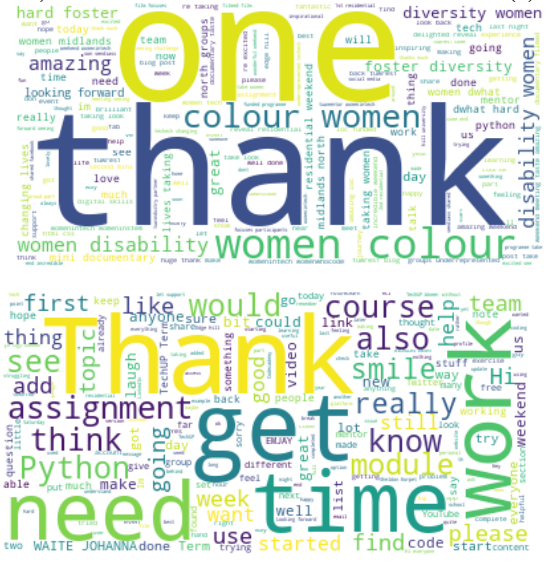

Fig. 2. Word cloud analysis of posts from Twitter (top) and Microsoft Teams (bottom).

For our experiment, we compare the output of running LDA with 2, 4, 6, 8 and 10 topics. We then have qualitatively evaluated the topic model outputs, to determine the extent to which the identified clusters, representing a topic, are semantically related to our datasets. Although measuring topic coherence is a common approach to determine the best number of topics to select, qualitative judgement to determine the performance of topic models have also proven useful as suggested in prior research [37]. We selected 10 topics because this value proved to reveal the most relevant themes based on our datasets.

TABLE I. VOCABULARY FOR EACH TOPIC GENERATED FROM LDA WHEN K = 10. THE VOCABULARIES FOR EACH TOPIC ARE RANKED ACCORDING TO THEIR WEIGHTS IN DESCENDING ORDER

Topics clusters	**Vocabularies**
Topic 1	*thank, look, realli, great, share, forward, techupwomen, mentor, help, home*
Topic 2	*tuw, tuwres, talk, video, got, watch, techupwomen, womenintech, residenti, inspir*
Topic 3	*love, residenti, weekend, ye, game, think, time, fab, happen, peopl*
Topic 4	*done, techupwomen, awesom, know, happi, python, assign, cours, techup, work*
Topic 5	*tuwres, techupwomen, see, today, challeng, film, weekend, readi, tuw, dream*
Topic 6	*excit, wait, techupwomen, wonder, tuwres, inspir, learn, week, tomorrow, tuw*
Topic 7	*tech, start, techupforwomen, day, fantast, women, techupwome, womenintech, absolut, techupwomen*
Topic 8	*take, check, women, chang, live, techupwomen, midland, look, womenintech, incred*
Topic 9	*techupwomen, tuw, tuwres, welcom, part, womenintech, meet, edg, hill, photo*
Topic 10	*amaz, go, proud, hope, programm, enjoy, opportun, brilliant, thought, women*

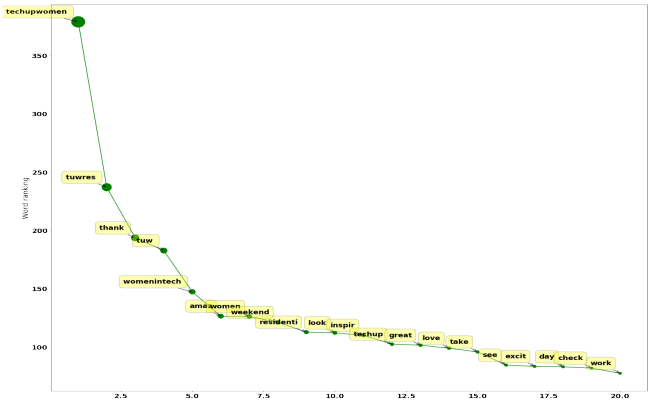

Fig. 3. Ranking of the top 20 terms on Twitter using term TF-IDF.

The topics generated from the LDA on Twitter are shown in Table 1. With unsupervised topic models, many topics contain words in each cluster that are similar and overlapping. A common theme across the topics is that they frequently contain the hashtags (e.g. "techupwomen", "womenintech", "tuwres") that are related to the programme. These hashtags give indications of the context of the topics that are identified. Examining the topic clusters, the vocabularies include words that provide the context of some of the posts on Twitter.

Terms such as (“proud”, “brilliant”, “opportune” - for opportunity) are words that might be observed to express positive sentiments towards the programme. More importantly, some topics also include words consistent with promoting the participation of women in technology. For example, Topic 8 contains “change”, “live”. Others are locality-dependent, such as “midland”. Some other topics are also associated to the topic learning, including “Python”, “assign”, “course”, “work”, “done”. Other topics also contain words that are consistent with special events, including weekend residentials (e.g. “weekend”, “people”, “happen”).

In this section, we present two visualisations of networks based on Twitter text. One is based on intersections between vertices in the network (Figure 4) and represents a compact temporal representation of the networking process. Specifically, we analyse the graph connections monthly between July - October. There are 852 vertices and 3269 edges in total. The red edges represent the connections between users in July, orange edges represent the connection in August, yellow edges represent the connections in September while the green edges represent the connections for October. The graph shows both the in-degree connections, the number of incoming edges to the vertex and out degree, the number of outgoing edges from the vertex [38]. The network graph is useful to visualise the engagements and flow of information during this period. Overall, the network graph is very dense and with significant amount of interactions between the Twitter users. In July and September, the networks are the densest and connect a larger number of vertices (see Table II). Both these periods are when the residentials were held. This is consistent with the previous trends, the programme residential is a unique characteristic that shows strong social relationship and active involvement by participants and other relevant actors.

TABLE II. OVERALL METRICS FOR NETWORK GRAPH PER MONTH

Month	*Edges*	*Vertices*
July	963	270
August	729	270
September	984	356
October	593	226

Another network analysis is shown based on measuring the eigenvalues for centrality across the network Figure 5 presents the visualisation of the top 10 vertices, based on the eigenvalue centrality measurement. The results show that the important players in the programme play different roles in the social network. The TechUP official Twitter account, *@TechUpWomen* appears in the centre of the network for all the months. Other vertices at the core of the network include *@compscidurham*, *@edgehill*, *@IoCoding* and *@durham_uni*. The most ranked vertices in terms of betweenness are organisations and individuals that are involved in the program. They play important roles in the network by serving as a hub for engagements and interactions on Twitter. Some other users such as user4, user3 and user1 also play important roles in linking external vertices to the core network.

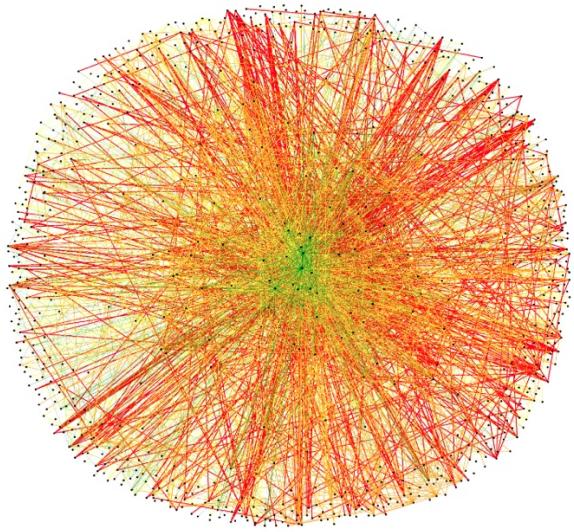

Fig. 4. Network Analysis Graph Visualisation with Twitter Data.

V. CONCLUSIONS

In this paper, we have introduced, for the first time, a digital inclusion programme specifically aimed at re-training women in underdeveloped areas - Midlands and North of England, where technical know-how is required by many companies. The programme has various novel and ambitious features, such as one-to-one mentoring (i.e., 100 mentors for 100 women, and an additional mentoring network), networking support (online and offline), residential weekends, close connection to industry (industry-driven programme), strong presence and support of under-represented communities (Black, Asian and Minority Ethnic (BAME), disabilities, dependents), diversity, feeding back to the community (women ambassadors), human-centric (enabling instead of teaching approach; motivation enhancing), guaranteed interviews at the end of the programme, free entry for the women. Thus, the paper responds to our second research question, RQ2, on how special intervention methods can support women transition into technology roles.

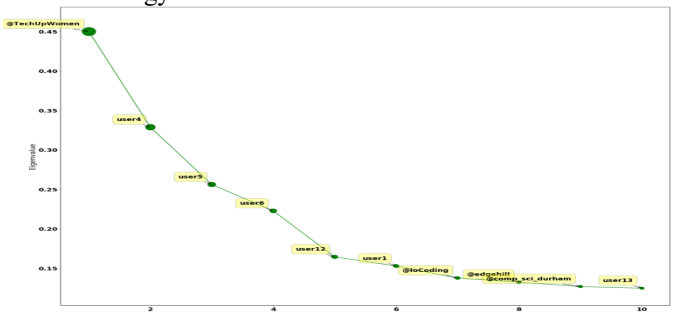

Fig. 5. Network Analysis Visualisation for Top 10 Nodes based on Eigenvalue Centrality

However, we offer more than this, by constructing a set of measurements, based on data analytics, that allow us to estimate the support offered, even during the programme itself, via nowcasting. The temporal representation of the volume of posts shows a clear correlation with actual events in the programme, and their effect can then be measured. The topics of discussion (on Twitter or Microsoft Teams) show the similarities and differences in focus during the course,

presented here a visual or numerical manner. Temporal evolution of trends is also a useful tool to estimate how the focus of participants changes, and, comparing his with the expected focus, to plan and provide timely interventions during nowcasting. Clustering can bring together clusters (topic groups) based on their similarity, and thus the different axes of opinions of participants can be estimated. Topic intersection representations and triangularisation techniques have shown the most popular topics based on several methods, and thus confirmed via different ways. Thus, we can confidently say that the TechUPWomen programme has been so far exciting, inspiring and great, amongst others, and that our participants are very thankful. Finally, network analysis can represent influential users and spread of discussions at a given moment in time, or over time. We can thus measure central participants in the discussions, and potentially nudge others to participate more actively. For further work, we shall apply sentiment analysis on the social media data, to have a clearer understanding of the exact nature of the sentiments expressed by our participants.

REFERENCES

[1] L. J. Sax, H. B. Zimmerman, J. M. Blaney, B. Toven-Lindsey, and K. J. Lehman, "Diversifying undergraduate computer science: The role of department chairs in promoting gender and racial diversity," *J. Women Minor. Sci. Eng.*, vol. 23, no. 2, pp. 101–119, 2017.

[2] S. Seibel and N. Veilleux, "Factors Influencing Women Entering the Software Development Field through Coding Bootcamps vs. Computer Science Bachelor's Degrees *," 2019.

[3] K. Singh, K. R. Allen, R. Scheckler, and L. Darlington, "Women in computer-related majors: A critical synthesis of research and theory from 1994 to 2005," *Review of Educational Research*, vol. 77, no. 4. pp. 500–533, Dec-2007.

[4] E. L. Wilder, L. A. Tabak, R. I. Pettigrew, and F. S. Collins, "Biomedical research: Strength from diversity," *Science*, vol. 342, no. 6160. p. 798, 2013.

[5] L. Waguespack, J. S. Babb, and D. Yates, "Triangulating Coding Bootcamps in IS Education: Bootleg Education or Disruptive Innovation?," 2018.

[6] F. Xiong and Y. Liu, "Opinion formation on social media: an empirical approach," *Chaos*, vol. 24, no. 1, p. 013130, Mar. 2014.

[7] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson, "Using Internet Searches for Influenza Surveillance," *Clin. Infect. Dis.*, vol. 47, no. 11, pp. 1443–1448, Dec. 2008.

[8] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.

[9] S. Baker and A. Fradkin, "What Drives Job Search? Evidence from Google Search Data," 2011.

[10] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS One*, vol. 6, no. 12, Dec. 2011.

[11] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. Google trends," *J. Forecast.*, vol. 30, no. 6, pp. 565–578, Sep. 2011.

[12] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 41, pp. 17486–17490, Oct. 2010.

[13] V. Lampos, T. Lansdall-Welfare, R. Araya, and N. Cristianini, "Analysing Mood Patterns in the United Kingdom through Twitter Content," 2013.

[14] H. Choi and H. Varian, "Predicting the Present with Google Trends," *Econ. Rec.*, vol. 88, no. SUPPL.1, pp. 2–9, Jun. 2012.

[15] V. Lampos and N. Cristianini, "Nowcasting Events from the Social Web with Statistical Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–22, 2012.

[16] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Effects of the recession on public mood in the UK," in *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*, 2012, pp. 1221–1226.

[17] H. K. Green, N. J. Andrews, G. Bickler, and R. G. Pebody, "Rapid estimation of excess mortality: Nowcasting during the heatwave alert in England and Wales in June 2011," *J. Epidemiol. Community Health*, vol. 66, no. 10, pp. 866–868, Oct. 2012.

[18] J. Ray and J. S. Brownstein, "SANDIA REPORT Nowcasting influenza outbreaks using open-source media reports."

[19] V. Lampos, "On voting intentions inference from Twitter content: a case study on UK 2010 General Election," Apr. 2012.

[20] L. Chen, K. S. M. Tozammel Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models," *Data Min. Knowl. Discov.*, vol. 30, no. 3, pp. 681–710, May 2016.

[21] D. Pruss *et al.*, "Zika discourse in the Americas: A multilingual topic analysis of Twitter," *PLoS One*, vol. 14, no. 5, p. e0216922, May 2019.

[22] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach," in *ACM International Conference Proceeding Series*, 2015, vol. 16-20-Marc, pp. 146–150.

[23] S. Ghosh *et al.*, "Temporal Topic Modeling to Assess Associations between News Trends and Infectious Disease Outbreaks," *Sci. Rep.*, vol. 7, 2017.

[24] M. J. Paul and M. Dredze, "Discovering Health Topics in Social Media Using Topic Models," *PLoS One*, vol. 9, no. 8, p. e103408, Aug. 2014.

[25] D. M. Blei, A. Y. Ng, and M. T. Jordan, "Latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2002, vol. 3, pp. 993–1022.

[26] A. E. Cano Basave, Y. He, and R. Xu, "Automatic labelling of topic models learned from Twitter by summarisation," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 2014, vol. 2, pp. 618–624.

[27] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *J. Inf. Sci.*, vol. 43, no. 1, pp. 88–102, Feb. 2017.

[28] D. Ediger *et al.*, "Massive social network analysis: Mining twitter for social good," in *Proceedings of the International Conference on Parallel Processing*, 2010, pp. 583–593.

[29] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977.

[30] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001.

[31] F. Liljeros, C. R. Edling, L. A. Nunes Amaral, H. E. Stanley, and Y. Åberg, "Social networks: The web of human sexual contacts," *Nature*, vol. 411, no. 6840, pp. 907–908, Jun. 2001.

[32] B. Güçlü, M. Á. Canela, and I. Alegre, "An Exploratory Analysis Using Co-Authorship Network," 2018, pp. 166–200.

[33] J. Uyheng and K. M. Carley, "Characterizing bot networks on twitter: An empirical analysis of contentious issues in the Asia-Pacific," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11549 LNCS, pp. 153–162.

[34] K. Kandasamy and P. Koroth, "An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques," in *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2014*, 2014.

[35] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," 2011.

[36] A. Hagberg, P. Swart, and D. Chult, "Exploring network structure, dynamics, and function using NetworkX," 2008.

[37] S. Muralidhara and M. J. Paul, "#Healthy Selfies: Exploration of Health Topics on Instagram," *JMIR Public Heal. Surveill.*, vol. 4, no. 2, p. e10150, 2018.

[38] J. Kim and M. Hastak, "Social network analysis: Characteristics of online social networks after a disaster," *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 86–96, 2018.